Researchers use deep learning to build automatic speech recognition system to help preserve the Seneca language

15 October 2018, by Michelle Cometa

Left to right, Ray Ptucha, computer engineering assistant professor, Robbie Jimerson, computer science doctoral student, both from RIT, and Emily Prud'hommeaux, assistant professor of computer science, are leading the NSF project to use artificial intelligence technology to preserve the Seneca language. Credit: A. Sue Weisler/RIT

A new research project at Rochester Institute of Technology will help ensure the endangered language of the Seneca Indian Nation will be preserved. Using deep learning, a form of artificial intelligence, RIT researchers are building an automatic speech recognition application to document and transcribe the traditional language of the Seneca people. The work is also intended to be a technological resource to preserve other rare or vanishing languages.

"The motivation for this is personal. The first step in the preservation and revitalization of our language is documentation of it," said Robert Jimerson (Seneca), a computing and information sciences doctoral student at RIT and member of the research team. He brought together tribal elders and close friends, all speakers of Seneca, to help produce audio and textual documentation of this Native American language spoken fluently by fewer than 50 individuals.

Like all languages, Seneca has different dialects. It also presents unique challenges because of its complex system for building new words, in which a whole sentence can be expressed in a single word.

Jimerson is able to bridge both the technology and the language.

"Under the hood, it is data. With many Native languages, you don't have that volume of data," he said, explaining that some languages, while spoken, may not have as many formal linguistical tools—dictionaries, grammatical materials or extensive classes for non-native speakers, similar to those for Spanish or Chinese. "One of the most expensive and time-consuming processes of documenting language is collecting and transcribing it. We are looking at taking deep networks and maybe changing the architecture, making some synthetic data to create more data, but how do you make this work in deep learning? How do you augment data you already have?"

That process of attaining data is being coordinated by a wide-ranging team that includes Jimerson; the project principal investigator Emily Prud'hommeaux, assistant professor of computer science at Boston College and research faculty in RIT's College of Liberal Arts; Ray Ptucha, assistant professor of computer engineering in RIT's Kate Gleason College of Engineering and an expert in deep learning systems and technologies; and Karen Michaelson, professor of linguistics, the State University of New York at Buffalo. The research team was awarded $181, 682 in funding over four years from the National Science Foundation for

"Collaborative Research: Deep learning speech recognition for document Seneca and other acutely under-resourced languages."

"This is an exciting project because it brings together people from so many disciplines and backgrounds, from engineering and computer science to linguistics and language pedagogy," said Prud'hommeaux. "In addition to enabling us to develop cutting edge technology, this project supports undergraduate and graduate students and engages members of an indigenous community that few people know is right here in western New York."

The researchers started the project in late June, bringing together the community members and linguists for data collection—acquiring and translating current and new, original recordings of Seneca conversations then converting data into textual output using deep learning models.

"What you are really trying to do is find that line between the new data you can get and the changing of the architecture of a network," Jimerson explained.

Since the summer, the team has just over 50 hours of recorded material with people working full time on the translations that include breaking down the language into individual phonetic symbols and using this information to begin training the models.

"We use a process called transfer learning which starts with a model trained with readily available English speech to get the basic, initial training for the system, then we'll re-train the neural networks and fine tune it toward the Seneca language. We're getting very good results," said Ptucha, who is an expert in deep learning systems and technologies. Deep learning technology consists of multiple layers of artificial neurons, organized in an increasingly abstract hierarchy. These architectures have produced state-of-the-art results on all types of pattern recognition problems including image and speech recognition applications.

"No one has really tried this before, training an automated speech recognition model on something as resource-constrained as Seneca. Robbie is the expert in transcribing Seneca and training the others on how to do this. He's a pretty rare guy, " said Ptucha,

This current project is a continuation of Jimerson's work to expand the language resources available to his community. In 2013, while he was a graduate student in RIT's Golisano College of Computing and Information Sciences, he developed an online Seneca language translation dictionary for the Seneca Language Revitalization Program. The project was funded by the Seneca Nation and awarded to RIT's Future Steward's Program.

Provided by Rochester Institute of Technology